A European Football Betting Model: Utilizing Poisson Distributions & Kelly Criterion to Maximize Profits

Abstract

Sports betting has experienced a rapid rise in popularity as accessibility and commercialization of daily fantasy and live betting has increased. As sports betting is legalized in different countries and states, we are presented with a new opportunity to create statistical models that we can utilize to predict outcomes of different sporting events that can then be used to find inefficiencies in sportsbooks. In this paper, we attempt to create a model for European Soccer and measure its performance against betting markets to understand if this model can be used to generate profits. The paper shows how we assigned teams an attacking and defensive rating based on their production against their league counterparts which we then utilized to create a Poisson distribution that determined the probability of each event (Win, Lose, Draw). To add another dimension to our model, we used an optimization technique known as Kelly Criterion to determine the optimal amount of money that should be bet on each match. This technique generates bet amounts while also creating a value (KCO Value) that acts as an accurate estimator of the risk associated with each match. By exploring the characteristics of this value, we were able to maximize the success of the model. After running the model for the 2018 and 2019 seasons across the five major European soccer leagues, we can safely say that our model was not only successful in predicting outcomes, but also in generating significant profit yields for a user. A profit percentage of 105.93% was yielded using this model, which implies that a user would pretty much double their money using this model. We also evaluate how the model performs in different leagues to understand which league characteristics benefit the model. The highest profit percentage was seen in the 2019 Premier League season with a profit percentage of 151%. The success of the model can not only help users generate significant profits, but it can also expose certain inefficiencies in the market.

Data and Methodology

The first step was to create a predictive model that can update as the season goes on. After every game week, the model adapted to take into consideration team trends and to improve accuracy. The model predicts the scoreline of a game using a Poisson Distribution[1]. A Poisson Distribution is used to predict the number of events within a certain time interval. In this case we will use the number of goals[2] as the events we want to predict, and the time interval being the length of the game. The formula is as follows, with x representing the number of events and λ representing the mean.

$$P(x) = \frac{e^{-\lambda} \times \lambda^{x}}{x!}$$

The missing item from the formula is λ. To calculate the mean value, we developed an attacking strength and defending strength for every team.

$$Attacking\ Strength = \frac{Goals\ Per\ Game\ For\ The\ Team}{Goals\ Per\ Game\ Across\ The\ League}$$

$$Defending\ Strength = \frac{Goals\ Allowed\ Per\ Game\ For\ The\ Team}{Goals\ Allowed\ Per\ Game\ Across\ The\ League}$$

These are formulas that have widely been used in the community. The attacking and defending strengths for each team change after every week as the average goals scored and allowed update. These have also been divided into home and away splits to further improve accuracy. However, these are not the final strengths that we use for each game. A team's lambda value for a game is calculated using the formula:

$$HomeGoalsPred = HomeAttackingStrength \times AwayDefendingStrength \times AverageLeagueHomeGoals$$

$$AwayGoalsPred = AwayAttackingStrength \times HomeDefendingStrength \times AverageLeagueAwayGoals$$

These prediction values are each team's respective λ value. This λ value considers the teams attacking strength and the opponents defending strength. The average league home and away goal variables represent the average home field advantage across an entire league which is derived from the total goals scored by home and away teams and used as a multiplier. Since we have a formula for the λ value each game we can use the Poisson formula mentioned above to calculate the probabilities for a home win, away win, and a draw. We used the formula to calculate the probability for every possible scoreline from 0 to 0 to 5 to 5. We can create a result matrix which contains the probabilities for every possible scoreline. The probability for each scoreline is calculated using the following formula:

$$Prob\ of\ Scoreline = P(HomeGoals) \times P(AwayGoals)$$

The probabilities for each possible outcome are calculated in the following manner:

$$HomeWin = \sum Probabilities\ of\ Scorelines\ with\ Home\ Team\ Winning$$

$$AwayWin = \sum Probabilities\ of\ Scorelines\ with\ Away\ Team\ Winning$$

$$Draw = \sum Probabilities\ of\ Scorelines\ ending\ with\ a\ Draw$$

To provide more clarity, we will use the Crystal Palace (H) vs West Brom (A) game as an example during the last game week of the 2018 Premier League season. Below is a calculation of both team's attacking and defending strengths along with our goal prediction for them:

$$CRY\ Att\ Strength = \frac{1.5}{1.5189} = 0.9875 \qquad WBA\ Att\ Strength = \frac{0.5556}{1.1486} = 0.4837$$

$$CRY\ Def\ Strength = \frac{1.5}{1.1486} = 1.3059 \qquad WBA\ Def\ Strength = \frac{1.3889}{1.5189} = 0.9144$$

$$CRY\ Goals\ Pred = 0.9875 \times 0.9144 \times 1.5189 = 1.3716$$

$$WBA\ Goals\ Pred = 0.4837 \times 1.3059 \times 1.1486 = 0.7255$$

Since we have our goal predictions for each team, we can use the Poisson distribution to calculate the probability of each scoreline. Below is a graph that depicts the results for this game:

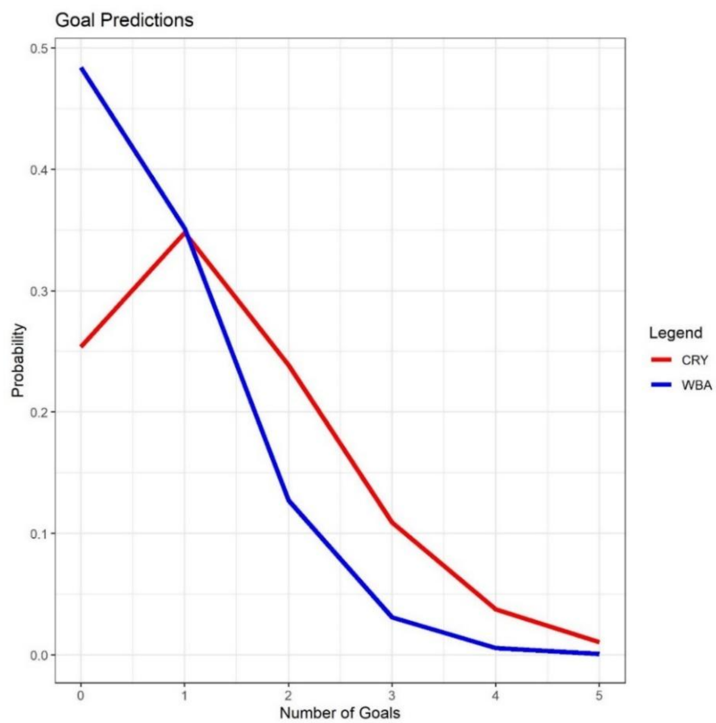

Figure 1: Probability density function of projected goals in sample match

The horizontal axis depicts the number of goals and the vertical axis depicts the probability of the team scoring that many goals. Based on our model, West Brom has the highest probability to score 0 goals and Crystal Palace has the highest probability to score 1 goal. The likelihood of each score line is calculated by multiplying the probabilities as discussed earlier. Based on the graph, we can see that the model believes the most likely scoreline is 1 – 0 and predicts that this has a 0.168 (0.348×0.4841) chance of occurring. To determine each outcome's probability, we can sum all the instances we obtain for the desired outcome. The results are below:

$$Crystal\ Palace\ Win = 0.5202 \qquad West\ Brom\ Win = 0.1977 \qquad Draw = 0.279$$

To ensure that the model had a baseline to create accurate values for a team's attacking and defending strength, we used the model for games after game week 13. This resulted in a total sample size of 2378 games. After developing the predictive model, the next step was deciding how the bets would be placed and distributed based on our outputs. There are many different bankroll management methods employed by bettors across the community. It is very common for bettors to use fixed units or percentages while others differ their bet amounts by how confident they are on their bets. We used a bankroll of $100 for each game so that each bet could be made independent of the success of another bet. To measure the accuracy of the predictive model better, we used the Kelly Criterion Model[3] to dictate the amount bet on each game.

The Kelly Criterion model is a form of probability theory most commonly used by investors that has made its way into the betting world. The goal of the model is to maximize our profit while accounting for the risk associated with a lost bet. This is done by maximizing the logarithm of the potential ending bankrolls after the bet is placed. Many online sources choose to simplify the math behind the Kelly Criterion Theory into a generic formula that looks like this:

$$\frac{O \times (P_w - P_l)}{O} = B$$

$$O = Odds$$

$$P_w = Probability\ of\ Bet\ Winning\ and\ P_l = Probability\ of\ Bet\ Losing,$$

$$B = Percentage\ of\ Bankroll\ to\ Bet$$

While this formula does involve some principles of the Kelly Criterion Theory, it does not reflect the theory itself. Also, a soccer match has three outcomes, slightly complicating the math of this simplified two outcome formula. Our goal is to simply maximize the logarithm of our potential bankroll according to the Kelly Criterion Theory.

Bet	Decimal Odds	Win Probability	Bet Amount
Home	1.8	52.02%	$0.00
Away	4.75	19.77%	$0.00
Draw	3.79	27.90%	$0.00

Outcome	Probability	Starting Bankroll	Wins	Losses	Ending Bankroll	Logarithm
Home Win	52.02%	$100.00	$0.00	$0.00	$100.00	2
Away Win	19.77%	$100.00	$0.00	$0.00	$100.00	2
Draw	27.90%	$100.00	$0.00	$0.00	$100.00	2

Objective 2.000000

Table 1: Example of the bet optimizer before running the solver add-in

Shown above is an example of how our optimization looked for the example discussed earlier. The inputs are the odds[4] and probabilities of each possible result as well as the starting bankroll. Using this information, we can calculate our ending bankroll for each possible event and then take the logarithm of that. Lastly, the "objective" is a weighted average of the logarithms and their associated probabilities. Once this information is set, we use Microsoft Excel's solver add-in to maximize the objective cell by changing the "Bet Amount" cells. A typical output can be seen below:

Bet	Decimal Odds	Win Probability	Bet Amount
Home	1.8	52.02%	$22.34
Away	4.75	19.77%	$1.65
Draw	3.79	27.90%	$0.00

Outcome	Probability	Starting Bankroll	Wins	Losses	ding Bankr	Logarithm
Home Win	52.02%	$100.00	$17.87	$1.65	$116.22	2.065286
Away Win	19.77%	$100.00	$6.19	$22.34	$83.85	1.923481
Draw	27.90%	$100.00	$0.00	$23.99	$76.01	1.880889

Objective 2.020917

Table 1: Example of our bet optimizer after running the solver add-in

Based on these results, the model predicts a bet of $22.34 on Crystal Palace to win and a bet of $1.65 on West Brom to win. The result of the game was 2 – 0 in favor of Crystal Palace, hence we would have profited $16.22 of this bet.

It is important to understand that since the logarithm value is derived from the predicted ending bankroll, a higher number indicates a larger potential profit. The final objective is the sum of each logarithm value multiplied by the probability of the outcome; therefore, the value is indicative of the amount of risk associated with each bet. Individual bets below the threshold of 2 may return a higher profit than games closer to 2.1, but there is a greater chance of winning consistently by betting on the higher log values. The last aspect of the KCO model that considers risk is when the maximized profit includes betting on both a win and a draw or a loss and a draw. The odds and our percentage model can combine in a way that the necessary procedure for reducing risk is to bet on multiple outcomes. There are also games in which the optimization reveals that no bets should be placed based on the odds and model predictions.

Model Evaluation

In this section, we will observe how successful the model was in terms of overall profit. It is important to note that the Kelly Criterion suggests different amounts to bet for each game based on the risk, hence we decided to use profit percentage as an indicator for the model's success in addition to profit.

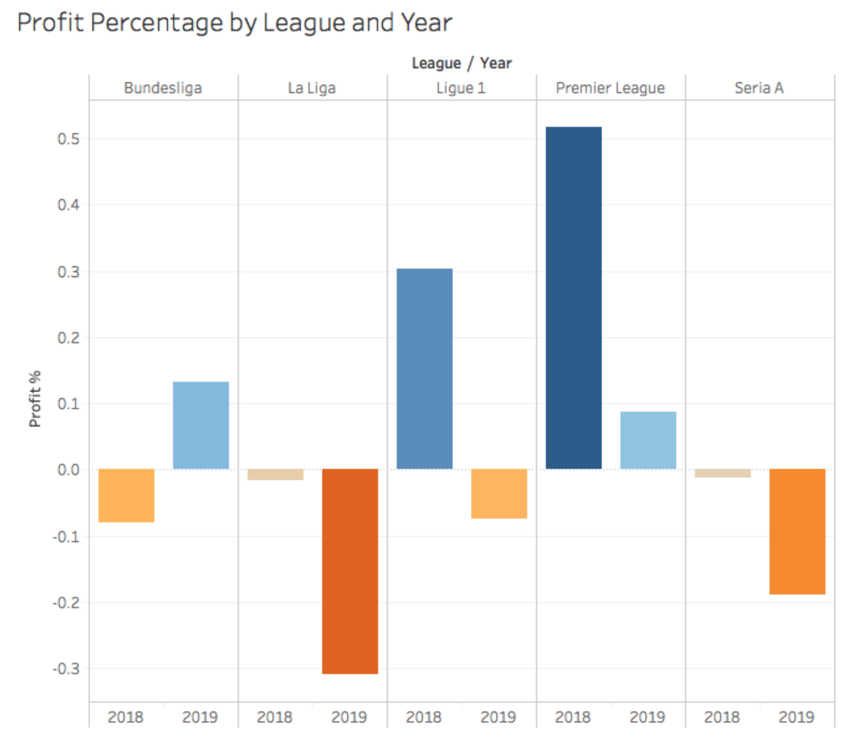

Figure 2: A graph showing the profits and losses for the model in each league

As pictured above, the Premier League and Ligue 1 in 2018 were highly successful using our model. Specifically, the Premier League in 2018 generated $4,966.72 at 52% profit while Ligue 1 had a total profit of $2,635.81 at 30%. While the overall results of the ten samples seasons were unsatisfactory, we were able to discover that the success rate and profit can be significantly increased when filtering out games with a KCO generated log value lower than 2.

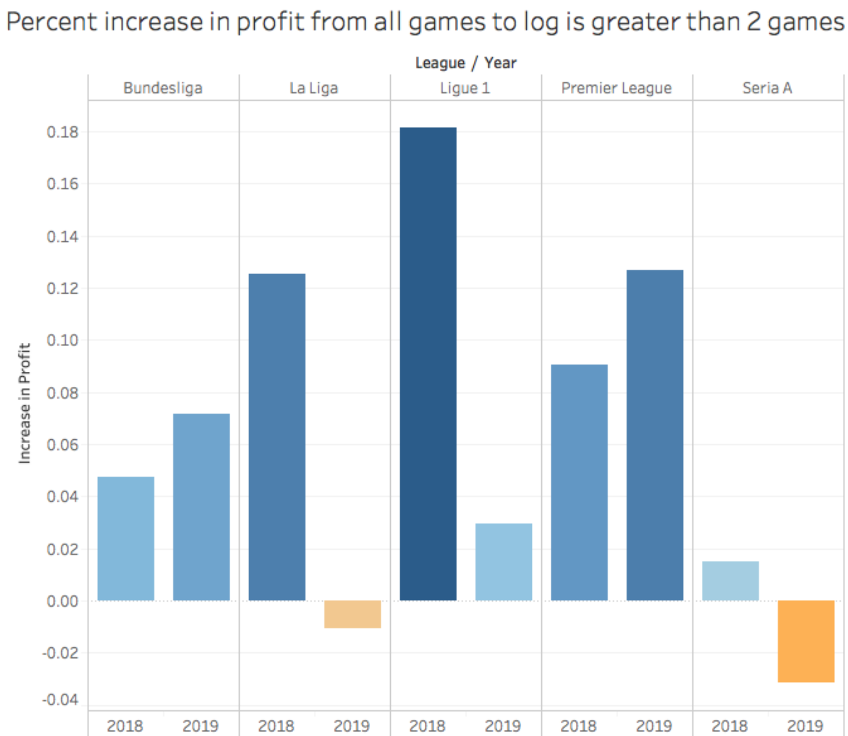

Figure 3: Results when betting only games with a KCO log value > 2

When only looking at games greater than 2, eight out of the ten leagues experience an increase in profit percentage. Ligue 1 in 2019, Bundesliga in 2018, and Serie A in 2018 all went from negative or breakeven to positive dollar profit when only betting on confident games. In other words, the amount of risk associated with higher logarithm values is less than when a game is at or below 2. We also noticed that games with a higher KCO value tend to have higher odds for a draw. In this instance, the model usually hedges[1] its bet by throwing a small amount on the draw, allowing for either a lower profit or a smaller loss if a draw occurs rather than the predicted team winning. This explains why a higher KCO value depicts a lower risk.

As we have now established that the KCO is an accurate estimator of risk. The profit percentage can be further increased when examining the Premier League, the league that fits our model best. As the confidence got closer to 2.1 in 2018 and 2019, the profit percentage gradually increased up to 141% and 153% respectively when only betting on the games our model was most confident in.

[1] "Hedging a bet is a strategy in which a bettor will place a second wager against the original bet when they're unsure that the outcome of a wager will be a win."

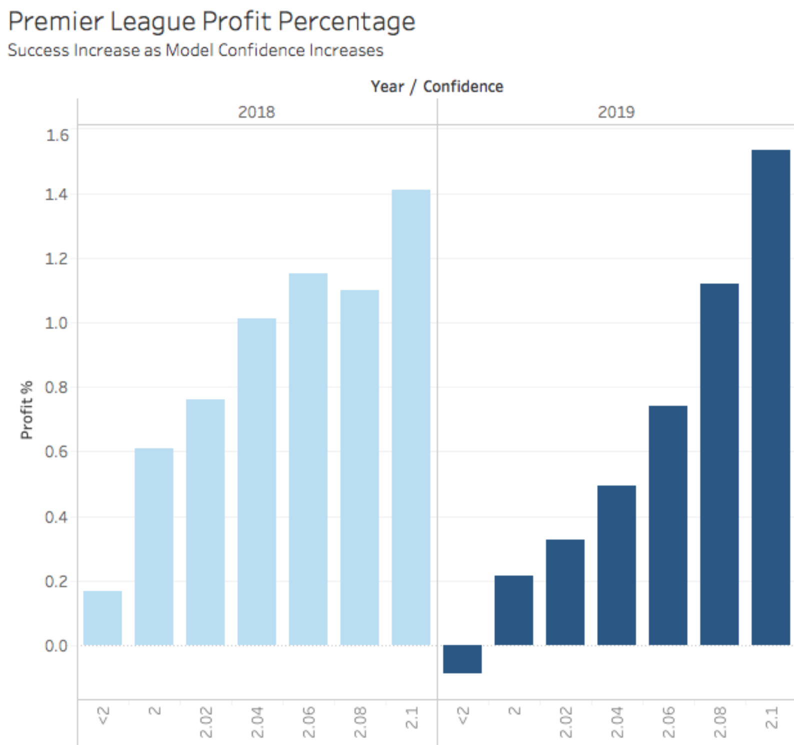

Figure 4: Profit percentage for the Premier League based on the model's log value

Here is a table that summarizes the profit percentages for each league when looking at games that have a log value of less than 2 and greater than 2.1:

	Log Value < 2		Log Value >2.1	
League	Profit Percentage	Amount Risked	Profit Percentage	Amount Risked
Premier League 2018	17%	$1,974.19	141%	$1,997.02
Premier League 2019	-9%	$2,820.04	153%	$769.87
Ligue 1 2018	-20%	$2,303.00	132%	$2,022.57
Ligue 1 2019	-14%	$1,801.95	-100%*	$129.44
La Liga 2018	-20%	$2,688.87	71%	$119.11
La Liga 2019	-28%	$1,405.41	-27%	$349.21
Serie A 2018	-4%	$2,816.95	20%	$845.76
Serie A 2019	-15%	$2,636.63	-38%	$229.17
Bundesliga 2018	-15%	$1,717.24	97%	$54.40
Bundesliga 2019	7%	$2,143.15	123%*	$80.50

Table 3: Tabular view comparing low confidence and high confidence bets

*There were 0 games above 2.1, the values represent values for games with a log value greater than 2.08

This table highlights the accuracy and validates the use of the KCO value as an estimator for risk, as we can see that almost every league performs significantly better with a log value greater than 2.1 The best performance of the model was seen with the two Premier League seasons which obtained over 140% profit in both years. If we were to use the model for each of these leagues over 2 years for games that had a log value greater than 2.1, we would have made $6,795.54 in profit with $6,381.11 being risked (Profit Percentage = 105.93%).

Possible Improvements

One of the aspects that our model did not touch on was how the power structure of the leagues affected the ability of our model to make money by predicting upsets. Inherently each of the 5 leagues have drastically different dynamics, and while the lambda value in the Poisson model added a league constant for home-field advantage, a future study regarding the impact parity has on betting odds and results would better indicate why the Premier League performs better with our model than Serie A. Another factor that could be further investigated would be to weigh the most recent weeks as having more impact on the probability of each outcome rather than using the teams' whole season equally. Lastly, if factors like injuries or more advanced metrics (xG, xGA, etc.) can be accounted for, the model may perform at a higher level.

Conclusion

When appropriately using past data to predict game outcome probabilities and then allocating funds efficiently to minimize risk it is clear that there is money to be made betting in European soccer. Utilizing our predictive model and the Kelly Criterion model, we found success in many leagues such as the Premier League, Ligue 1, and Bundesliga. However, there was still a sizable amount of loss being incurred. This led to the filtering of our KCO log values in each respective league. We found the most success and greatest profit for all games where the KCO log value was close to 2.1 with a profit percentage increasing up to 153% in the Premier League. As discussed above, there are always improvements that can be made with our model, including different methods following each league's dynamics. We hope to expand our model and continue to grow those profit percentages, finding the most accurate and profitable way to predict European soccer match outcomes.

References

[1] *How to calculate Poisson distribution for football betting.* (n.d.). Retrieved 2020, from https://help.smarkets.com/hc/en-gb/articles/115001457989-How-to-calculate-Poisson-distribution-for-football-betting

[2] *List of Competitions*. (n.d.). Retrieved 2020, from https://fbref.com/en/comps/

[3] *The real Kelly Criterion: A critical analysis of the popular staking method*. (2017). Retrieved 2020, from https://www.pinnacle.com/en/betting-articles/Betting-Strategy/the-real-kelly-criterion/HZKJTFCB3KNYN9CJ

[4] *Football Results, Statistics & Soccer Betting Odds Data*. (n.d.). Retrieved 2020, from https://www.football-data.co.uk/data.php

[Footnote 1] *How To Hedge A Bet: Sports Betting Strategy Explained*. (n.d.). Retrieved 2020, from https://www.thelines.com/betting/hedge